Interview: Artificial Intelligence, Natural Intelligence and Evolution

Aaron Sloman, interviewed by Adam A. Ford

http://www.cs.bham.ac.uk/~axs

http://www.youtube.com/user/TheRationalFuture

March 2014

Abstract

This is a transcript of an interview, responding to questions by Adam Ford, introducing some of my most recent research. Adam interviewed me at the AGI "Winter Intelligence" conference at St Anne's College, Oxford, December 2012 http://agi-conf.org/2012/. He posed a number of questions to direct the interview, as indicated by the main section headings of this paper. Most of the questions, about the nature of AI, relationships between evolution and intelligence, and limitations in current AI were answered at some length, without interruption. However, in the final section the style changed to a dialogue concerning the future of AI/AGI. The difference is indicated by a change in format. The 57 minute video is available online at http://www.youtube.com/watch?v=iuH8dC7Snno. A draft (surprisingly accurate) transcript of the video was kindly produced (spontaneously) by Dylan Holmes at http://aurellem.org/ including inferred section headings! It was later revised and slightly extended, by me here http://www.cs.bham.ac.uk/research/projects/cogaff/movies/transcript-interview.html Jeremy Wyatt and Dean Petters then invited me to produce a version to be included with papers prepared for this "birthday" workshop on *"From Animals to Robots and Back: reflections on hard problems in the study of cognition"* http://www.cs.bham.ac.uk/~jlw/symposium_2011/book.html including an overview by Margaret Boden, who was not able to attend. However the fact that the transcript had already been made available under the Creative Commons "attribution" license made it unacceptable to the publishers.[1]

[1] An earlier version of the transcript was copied to H+ Magazine by Adam Ford in Sept 2013. Some portions of the interview have been re-written to improve clarity, without altering the original intended content, whereas in other places I have included additional material. A separate 2hr 30min video, by Adam Ford, recording Aaron Sloman's tutorial introduction to the Meta-Morphogenesis project the day after this interview was recorded, is available at http://www.youtube.com/watch?v=BNul52kFI74: Meta-Morphogenesis - How a Planet can produce Minds, Mathematics and Music. A high level overview, under continuous development, of the Meta-Morphogenesis project sketched in the interview is online at http://www.cs.bham.ac.uk/research/projects/cogaff/misc/meta-morphogenesis.html

Contents

1 Introduction - background, from mathematics through philosophy to AI

This section is mainly a response to a request to introduce myself and explain how I got into AI. It now has a little more detail than the recorded interview.

1.1 Evolving from doing mathematics to philosophy and AI

My first degree, many years ago in Cape Town, was in Physics and Mathematics, after which I intended to go on to do research in mathematics. However, while a student in Cape Town I had started attending philosophical meetings, where I met a philosophy graduate student, Adam Small, who recommended that I read Bertrand Russell's *History of Western Philosophy*. He was a Christian and I an atheist, but we got on very well. When I arrived in Oxford to study mathematics I continued my philosophical education, first by attending student philosophy societies and then by joining philosophy graduate seminars, where I got to know many excellent philosophers (several now famous).

The philosophers I met made claims about the nature of mathematics that I thought were seriously mistaken, and my arguments with them eventually triggered a decision to switch to a philosophy DPhil, defending Kant's view of mathematical knowledge. My supervisor was David Pears, although I was also supervised for a while by Hao Wang and later by Michael Dummett. My "personal tutor" in Balliol College, R.M. Hare, informally taught me meta-ethics. The DPhil thesis was completed in 1962[2], though for several years thereafter I felt that my defense of Kant needed improvement.

I then became a philosophy lecturer, first at Hull University, then, from 1964, at Sussex University. In 1969, I met Max Clowes[3] a leading AI vision researcher who introduced me to AI and taught me a great deal. I learned to program in several languages, of which my favourite was POP-2, the Edinburgh University AI Language, then later Pop-11[4]. Thanks to Max's encouragement, I wrote my first AI paper for the 2nd International Joint Conference in AI (IJCAI) in London 1971, attacking logicist AI as omitting some important forms of representation and reasoning. The paper was later reprinted in various places (Sloman, 1971). On the strength of that I was invited by Bernard Meltzer to spend the 1972-3 academic year in his Department at Edinburgh, where I met many outstanding AI researches and had my brain re-wired.

I soon became convinced that the best way to make progress in various areas of philosophy — including philosophy of mathematics, philosophy of mind, philosophy of language and other areas — was to try to design and test working fragments of a mind and perhaps eventually put them all together.

Later, after returning to Sussex, in collaboration with Max Clowes, Margaret Boden whom you've interviewed previously[5], and others, I helped to develop an interdisciplinary undergraduate degree combining AI with other subjects, and also began to do research in AI, which I thought

[2] http://www.cs.bham.ac.uk/research/projects/cogaff/sloman-1962/

[3] http://www.cs.bham.ac.uk/research/projects/cogaff/sloman-clowestribute.html

[4] http://en.wikipedia.org/wiki/POP-11

[5] http://www.youtube.com/watch?v=5dEXIOiAsaw

of as doing philosophy, primarily. For the next 17 years I worked on vision, architectures for intelligent systems, emotions and various other topics.

In 1991 I moved to the School of Computer Science in the University of Birmingham, where the work continued, in collaboration with students and colleagues, and for a while expanded into robotics.

I have been officially retired since 2002, but I am not interested in golf or gardening, so I just go on doing full time research, and so far my department seems to be happy to provide space and resources so that I can continue meeting bright people, while trying to learn and make progress on some hard problems.

1.2 AI is hard, in part because there are tempting non-problems.

One of the things I have learnt and understood more and more over the many years — forty years or so since I first encountered AI — is how hard the problems are. In part that's because it's very often tempting to *think* a problem is something different from what it actually is. When that happens, people design solutions to the non-problems, and I think of most of my work now as just helping to clarify what the problems are: what is it that we're trying to explain — and I hope this will lead into what you wanted to talk about.

I now think that one of the ways of getting a deep understanding of human intelligence is to find out what were the problems that biological evolution solved, because we are a product of *many* solutions to *many* problems, and if we just try to go in and work out what the whole system is doing, we may get it all wrong — badly wrong!

On the other hand if we simply focus on attempting to understand and model a particular aspect of human minds, e.g. vision, linguistic communication, planning, problem solving, or the roles of motives, emotions, moods and other types of affect, then we risk oversimplifying because each of those is connected with many other aspects and models that ignore the connections may not "scale out" when a richer set of requirements is considered.

Some researchers now hope to build a full working model of a brain by attempting to discover all the structures and connections in brains at a level of detail that allows components to be modelled electronically then assembled in a way that mirrors connections between components in brains. However, if much of the functionality of brains supports virtual machinery whose mapping onto physical structures and processes changes dynamically in very complex ways (as happens in multi-functional computers) then it is likely that important aspects of brain function will not be captured in the models built bottom up by attempting to match physical brain functions.

One way to avoid such traps, is to try to identify the many different transitions in information processing produced by biological evolution, developmental processes and learning, ever since the earliest organisms, independently of whether the functions are implemented directly in physical mechanisms or in virtual machinery. I call that The Meta-Morphogenesis (M-M) project, inspired in part by thinking about how to combine the ideas in Turing's early work on discrete computing machinery (Turing, 1936) with his later work on the chemical basis of Morphogenesis (Turing, 1952). The M-M project was first described in (Sloman, 2013) triggered by an invitation to contribute to (Cooper & Leeuwen, 2013). Some of the key ideas are presented in the next few sections, with further references in Section 10.

The "Meta" part of the name is a reminder that the mechanisms of change (including changes in mechanisms of natural selection, reproduction, development, and learning) can themselves produce new mechanisms of change, which would have been obvious to Turing, though it was not part of his paper on Chemical Morphogenesis.

2 What problems of intelligence did evolution solve?

2.1 Intelligence consists of solutions to many evolutionary problems; no single development (e.g. communication) was key to human-level intelligence.

First we need to challenge the assumption that we are the dominant type of organism (implied in the wording of the question not recorded in the video). I know it looks as if we are, but if you count biomass, if you count number of species, and if you count number of individuals, the dominant types of organism are microbes. Collectively they are dominant in those dimensions, and furthermore we are largely composed of microbes, without which we could not survive.

Humans are good at some things, e.g. doing mathematics, not so good at others, e.g. surviving without clothing in a blizzard. What allowed them to grow so (disastrously) numerous and powerful was a collection of different developments. Some people think that it was human language that made the crucial difference, where by "human language", they mean human communication in (spoken) words.

I suspect however, that that was a later development from what must have started as the use of richly structured *internal* forms of representation, which exist in nest-building birds, in pre-verbal children, and in hunting mammals, for example, because you can't take in information about a complex structured environment in which things can change, where you may have to be able to work out what's possible and what isn't possible, without having some way of representing the components of the environment, their relationships, the kinds of things they can and can't do, the kinds of things you might or might not be able to do, the consequences of various possibilities, and the constraints on what can occur or what can be done — and *that* kind of capability needs a powerful internal language, as argued in (Sloman, 1978, 1979; Sloman & Chappell, 2007; Sloman, 2008).

Some of us at Birmingham have been referring to the internal forms of representation as "generalized languages" (GLs) because many academics object (foolishly in my view) to using the term "language" to refer to something that isn't used for communication, and need not be used intentionally. But not only humans, but many other animals, have abilities to perceive complex and structured environments, and do things to their environment to make them more friendly to themselves, which depend on being able to represent possible futures and possible actions and to work out what's a good thing to do. This requires use of internal languages: GLs. The ability to deal with novel situations requires those GLs to support novel information structures, with some form of compositional semantics – usually thought of as unique to human external languages. I suggest that human sign languages with their rich mixture of discrete and continuous forms of representation will provide more powerful clues to the nature of GLs, and varieties of forms of syntax and semantics, than human spoken languages.

Nest-building achievements in corvids for instance (crows, magpies, rooks, and others) are far beyond what current robots can do. In fact I think most humans would be challenged if they had to go and find a collection of twigs, one at a time, maybe bring them with just one hand, or held in the mouth, and assemble them into a structure that is shaped like a nest, and is fairly rigid, so that you could trust your eggs in them when wind blows. But those birds have been doing it for a very long time. They are not our evolutionary ancestors, but they're an indication, an example, of what must have evolved in order to provide control over the environment in *other* species. It can be argued that those nest building processes would be impossible without the use of rich structured forms of representation of what is in the environment and possible ways of changing the environment, along with their consequences. (This would not apply too all nest building, for instance pushing bits of grass together.) One striking observation is that a crow will perform a complex task in several different ways on different occasions, all of which work, and without any random, trial-and-error, behaviours.

Insert: Movies[6] and photos[7] available at the University of Oxford Behavioural Ecology Research Group provide examples of what New Caledonian crows can do.

2.2 Speculation about how communication might have evolved from internal languages.

So I think hunting mammals, fruit-picking mammals, mammals that can rearrange parts of the environment, provide shelters, and so on, also needed to have ways of representing possible futures, not just what's there in the environment. I think at a later stage, that developed into a form of communication — or rather the *internal* forms of representation became usable as a basis for providing content to be communicated. That happened, I think, initially through performing actions that expressed intentions as a side-effect, which might have led to recognition of situations where an action (for instance, moving some large object) could be performed more easily, or more successfully, or more accurately if it was done collaboratively. So someone who had worked out what to do might start doing it, and then a conspecific might be able to work out what the intention is, because that person has the *same* forms of representation and can build theories about what's going on (in the actor's mind), and might then be able to help.

You can imagine that if that started happening often (a lot of collaboration based on inferred intentions and plans) then sometimes the inferences might be obscure and difficult, so the *actions* might be enhanced to provide signals as to what the intention is, and what the best way is to help, and so on.

Insert: (So actions enhanced to provide communication during collaboration may have been precursors to separately signed communications.)

So, this is all hand-waving and wild speculation, but I think it's consistent with a large collection of facts which one can look at — and find if one looks for them, but facts which one won't notice if one doesn't look for them, For instance, facts about the way children who can't yet talk, communicate, and the things they'll do, like going to the mother and turning her face to point in the direction where the child wants her to look and so on; that's an extreme version of action indicating intention.

Insert: A slide presentation elaborating this idea is here.[8]

Anyway. That's a very long roundabout answer to one conjecture that the use of communicative language is what gave humans their unique power to create and destroy and whatever, and I

[6] See `http://users.ox.ac.uk/.kgroup/tools/movies.shtml`

[7] See `http://users.ox.ac.uk/.kgroup/tools/crow.photos.shtml`

[8] See `http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#glang`

am saying that if by that you mean *communicative* language, then there was something before that which was *non*-communicative language, and I suspect that non-communicative languages continue to play a deep role in *all* human perception, in mathematical and scientific reasoning, and problem solving — and we don't understand very much about it.

I am sure there's a lot more to be said about the development of different kinds of senses, the development of brain structures and mechanisms to support all that, but perhaps I have droned on long enough on that question.

3 How do language and internal states relate to AI?

3.1 There are still major discrepancies between robots and animals

Most of the human and animal capabilities that I have been referring to are not yet to be found in current robots or computing systems, and I think there are two reasons for that: one is that it's intrinsically very difficult. In particular, it may turn out that the forms of information processing that one can implement on digital computers as we currently know them may not be as well suited to performing some of these tasks as other kinds of computing about which we don't know so much. For example, I think there may be important special features about *chemical* computers which we may talk about later.

3.2 In AI, false assumptions can lead investigators astray.

So, one of the problems is that the tasks are hard, but there's a deeper problem as to why AI hasn't made a great deal of progress on the problems that I am talking about, and that is that most AI researchers assume things — and not just AI researchers, but also philosophers, psychologists, and people studying animal behavior — they often make assumptions about what it is that animals or humans do. For instance they make assumptions about what vision is for, or assumptions about what motivation is and how motivation works, or assumptions about how learning works, and then AI people try to model or build systems that perform those assumed functions. So if you get the *functions* wrong, then even if you implement some of the functions that you're trying to implement, they won't necessarily perform the tasks that the initial objective was to imitate, for instance the tasks that humans, and nest-building birds, and monkeys and so on can perform.

3.3 Example: Vision is not just about finding surfaces, but about finding affordances.

I'll give you an example, to illustrate false assumptions. It is often assumed that the function of vision in humans (and in other animals with good eyesight and so on) is to take in optical information that hits the retina, usually forming constantly changing patterns of illumination, where there are sensory receptors that detect those patterns, and then somehow from that information (plus maybe other information gained from head movement or from comparisons between two eyes) to work out what there was in the environment that produced those patterns. And that is often taken to mean "Where were the surfaces off which the light bounced before it came to me?". So you essentially think of the task of the visual system as being to reverse the image formation process: so the 3D structure's there, the lens causes the image to form on the retina, and then the brain tries to build a model of that 3D structure out there.

This sort of theory is often attributed to David Marr (Marr, 1982). Compare the conjecture about brains as building models of the environment, in the final chapter of (Craik, 1943).

That's a very plausible theory about vision, and it may be that that's a *subset* of what human vision does, but, as James Gibson pointed out (J. Gibson, 1966; J. J. Gibson, 1979), that that kind of thing is not necessarily going to be very useful for an organism, and it is very unlikely that that's the main function of perception in general, namely to produce some physical description of what's out there.

I think there are far more kinds of "affordance" than Gibson noticed, and have presented some ideas about how to extend and generalise his work in a slide presentation (Sloman, 2011).

What does an animal *need*? It needs to know at least what it can do, what it can't do, what the consequences of its actions will be and so on. Gibson introduced the word "affordance": from his point of view, the functions of vision, or more generally perception, are to inform the organism of what the *affordances* are for action, where that would mean what the animal, *given* its morphology (what it can do with its mouth, its limbs, and so on, and the ways it can move) what it can do, what its needs are, what the obstacles are, and how the environment supports or obstructs those possible actions.

And that collection of required information structures is very different from the answer to "where are all the surfaces?". Even if you know where all the surfaces are, *deriving the affordances* would still be a major task.

So, if you think of the perceptual system as primarily (for biological organisms) being devices that provide information about affordances and so on, then the tasks look very different. And I think most of the people working on.., doing research on.., computer vision in robots, haven't taken all that on board, so they're trying to get machines to do things which, even if they were successful, would not make the robots very intelligent (and in fact, even the tasks they're trying to get robots to do are not really easy to do, and they don't succeed very well — although, there is progress: I shouldn't disparage it too much!)

3.4 Online and offline intelligence

It gets more complex as animals get more sophisticated. It is useful to make a distinction between *online* intelligence and *offline* intelligence. For example, if I want to pick something up — like this leaf (picks up a leaf from the table) — I was able to select it from all the others in there, and while moving my hand towards it, I was able to guide its trajectory, making sure it was going roughly in the right direction, as opposed to going out there, where I wouldn't have been able to pick it up. These two fingers ended up with a portion of the leaf between them, so that I was able to tell when I was ready to do the grasping (grasps the leaf between two fingers) and at that point, I clamped my fingers and then I could pick up the leaf.

That's an example of *online* intelligence: during the performance of an action (from the stage where it is initiated, and during the intermediate stages, until the action is completed) I am taking in information relevant to controlling all those stages, and that relevant information keeps changing. That means I need stores of transient information which gets discarded almost immediately and replaced, or modified. That's online intelligence. And there are many forms; that's just one example. James Gibson discussed many examples which I won't try to replicate now.

But in offline intelligence, you're not necessarily actually *performing* the actions when you're using your intelligence; you're thinking about *possible* actions. So, for instance, I could think about how fast or by what route I would get back to the lecture room if I wanted to get to the next talk or something. And I know where the door is, roughly speaking, and I know roughly which route I would take, when I go out, I should go to the left rather than to the right, because I have stored information about where the spaces are, where the buildings are, where the door was that we came out — but in using that information to think about that route, I am not actually performing the action. I am not even *simulating* it in detail: the precise details of direction and speed and when to clamp my fingers, or when to contract my leg muscles when walking, are all irrelevant to thinking about a good route, or thinking about the potential things that might happen on the way. Or what would be a good place to meet someone who I think frequents a particular bar or something. I don't necessarily have to work out exactly *where* the person is going to stand, or from what angle I would recognize him, and so on.

Offline intelligence – became not just a human competence. I think there are other animals that have aspects of it: Squirrels are very impressive. Gray squirrels at any rate, as you watch them defeating "squirrel-proof" bird feeders, seem to have a lot of that *offline intelligence*, as well as the online intelligence used when they eventually perform the action they've worked out would get them to the nuts.

I think that what happened during our evolution was that mechanisms for acquiring and processing and storing and manipulating information that was more and more remote from the performance of actions, developed. An example is taking in information about where locations are that you might need to go to infrequently: There's a store of a particular type of material that's good for building or putting on roofs of houses (or something); out around there in some direction. There's a good place to get water somewhere in another direction. There are people that you'd like to go and visit in another place, and so on.

So taking in information about an extended environment and building it into a structure that you can make use of for different purposes is another example of offline intelligence. And when we do that, we sometimes use only our brains, but in modern times, we also learned how to make maps on paper and walls and so on. And it is not clear whether the stuff inside our heads has the same structures as the maps we make on paper: the maps on paper have a different function; they may be used to communicate with others, or meant for *looking* at, whereas the stuff in your head you don't *look* at; you use it in some other way.

So, what I am getting at is that there's a great deal of human intelligence (and animal intelligence) which is involved in what's possible in the future, what exists in distant places, what might have happened in the past (sometimes you need to know why something is as it is, because that might be relevant to what you should or shouldn't do in the future, and so on), and I think there was something about human evolution that extended that offline intelligence way beyond that of other animals. And I don't think it was *just* human language, (but human language had something to do with it) but I think there was something else that came earlier than language (used for communication) which involves the ability to use your offline intelligence to discover something that has a rich mathematical structure.

3.5 Example: Even toddlers use sophisticated geometric knowledge

I'll give a simple example. If you look through a gap, you can see something that's on the other side of the gap. Now, you *might* see what you want to see, or you might see only part of it. If you want to see more of it, which way would you move?

Well, you could either move *sideways*, and see through the gap, and see roughly the same amount but a different part of it, if it is a room or whatever, or you could move *towards* the gap and then your view will widen as you approach the gap. Now, there's a bit of mathematics in there, insofar as you are implicitly assuming that information travels in straight lines, and, as you go closer to a gap, the straight lines that you can draw from where you are through the gap, widen as you approach that gap. Now, there's a kind of theorem of Euclidean geometry in there which I am not going to try to state very precisely (and as far as I know, wasn't stated explicitly in Euclidean geometry) but it is something every human toddler learns. (Maybe some other animals also know this?)

But there are many more things: more actions to perform, to get you more information about things, actions to perform to conceal information from other people, actions that will enable you, to operate, to act on, a rigid object in one place in order to produce an effect on another place.

So, there's a lot of stuff that involves lines and rotations and angles and speeds, and so on, that I think humans (and, to a lesser extent, other animals) developed the ability to think about in a generic way (using a combination of biological evolution and individual learning).

That meant that you could take the generalizations out of the particular contexts and then re-use them in new contexts in ways that I think are not yet represented at all in AI and in theories of human learning in any explicit way, although some people are trying to study learning of mathematics.

Insert: There has been a vast amount of research on how to give robots the ability to accumulate observational evidence and derive useful high probability generalisations. But the re-usable mathematical generalisations do not take the form of high probability generalisations based on large amounts of evidence: mathematical reasoning about geometric relationships is not concerned with probabilities, any more than theorems of arithmetic such as “3+5=8” or “There are infinitely many prime numbers” summarise statistical evidence.

4 Animal intelligence

4.1 The priority is *cataloguing* what competences have evolved, not ranking them.

I wasn't going to challenge the claim that humans can do more sophisticated forms of tracking, just to mention that there are some things that other animals can do which are in some ways comparable, and some ways superior to things that humans can do. In particular, there are species of birds and also, I think, some rodents, e.g. squirrels, and others that can hide nuts and remember where they've hidden them, and go back to them. And there have been tests which show that some birds are able to hide tens — eighty or more — nuts, and to remember which ones have been taken, which ones haven't, and so on. I suspect most humans can't do that. I wouldn't want to say categorically that we couldn't, because humans are very varied, and also a few people can develop particular competences through training. But it is certainly not something I can do (at present).

4.2 AI can be used to test philosophical theories

But I also would like to say that I am not myself particularly interested in trying to align animal intelligences according to any kind of scale of superiority; I am just trying to understand what it was that biological evolution produced, and how it works, and I am interested in AI *mainly* because I think that when one comes up with theories about how these things work, one needs to have some way of testing the theory.

And AI provides ways of implementing and testing theories that were not previously available: Immanuel Kant, e.g. in his *Critique of Pure Reason* (1781), was trying to come up with theories about how minds work, but he didn't have any kind of a mechanism that he could build to test his theory about the nature of mathematical knowledge, for instance, or how concepts were developed from babyhood onward. Whereas now, if we do develop a theory, we have a criterion of adequacy, namely it should be precise enough and rich enough and detailed to enable a model to be built. And then we can see if it works.

If it works, it doesn't mean we've proved that the theory is correct; it just shows it is a candidate. And if it doesn't work, then it is not a candidate as it stands; it would need to be modified in some way.

5 Is Artificial General Intelligence (AGI) feasible?

5.1 It is misleading to compare the brain and its neurons to a computer made of transistors

I think there's a lot of optimism based on false clues: for example, one of the false clues is to count the number of neurons in the brain, and then talk about the number of transistors you can fit into a computer or something, and then compare them. This comparison may be undermined by finding out more about how synapses work for example. I once heard a lecture in which a neuroscientist claimed that a typical synapse in the human brain has computational power comparable to the Internet a few years ago, because of the number of different molecules that are doing things, the variety of types of things that are being done in those molecular interactions, and the speed at which they happen, if you somehow count up the number of operations per second or something, then you get these comparison figures.

5.2 For example, brains may rely heavily on chemical information processing

Now even if the details aren't right, there may just be a lot of information processing that's going on in brains at the *molecular* level, not the neural level. Then, if that's the case, the processing units will be orders of magnitude larger in number than the number of neurons. And it is certainly the case that all the original biological forms of information processing were chemical; there weren't brains around, and still aren't in most microbes. And even when humans grow their brains, the process of starting from a fertilized egg and producing this rich and complex structure is, for much of the time, under the control of chemical computations, chemical information processing — of course constrained by sources of physical materials and energy also

So it would seem very strange if all that capability was something thrown away when you've got a brain, and all the information processing, the challenges that were handled in making a brain, were totally disconnected from the mechanisms used in a complete functioning brain. This is hand-waving on my part; I am just saying that we *may*, in future, learn that what brains do is not what we think they do, and that problems of replicating them are not what we think they are, and that our numerical estimates of time scales, numbers of components, needed, and so on could be wildly wrong.

5.3 Brain algorithms may be optimized for certain kinds of information processing other than bit manipulations

But there is another basis for skepticism about the imminence of AGI, namely doubts about how well we understand what *the problems* are, as opposed to the solutions. I think there are many people who try to formalize the problems of designing an intelligent system in terms of streams of information thought of as bit streams, or collections of bit streams, and they think of the problems of intelligence as being the construction or detection of patterns in those streams, and perhaps not just detection of patterns, but detection of patterns that are usable for sending *out* bit-streams to control motors, and so on, in order to achieve various goals, possibly defined in terms of desired input bit-streams.

That way of conceptualizing the problem may lead on the one hand to oversimplification, so that the things that *would* be achieved, if those goals were achieved, may be much simpler, and in some ways inadequate for replication or the matching of human intelligence — or, for that matter, squirrel intelligence — but in another way, it may also make the problems harder: it may be that some of the kinds of things that biological evolution has achieved can't be done that way, but can be done in different ways.

One of the ways that might turn out to be the case is not because it is impossible *in principle* to do some of the information processing on artificial computers, based on transistors and other bit-manipulating mechanisms; but it may just be that the *computational complexities* of solving problems, i.e. the complexities of processes or finding solutions to complex problems using bit manipulations, are much greater, and therefore you might need a much larger universe than we have available in order to do things – than if the underlying mechanisms were different.

Other *non bit-manipulating* information processing mechanisms might be better tailored to particular sorts of computation, as can be illustrated by examples.

5.4 Example: find the shortest path by dangling strings

There's a very well-known example, which is finding the shortest route if you've got a collection of roads, and they may be curved roads, and lots of tangled routes from A to B to C, and so on. If you start at A and you want to get to Z, a place somewhere on that map, the process of finding the shortest route will involve searching through all these different possibilities and rejecting some that are longer than others and so on. But if you make a model of that map out of strings, where the strings are all laid out on the maps and so have the lengths of the routes, then, if you hold the two knots in the network of strings which correspond to the start point and the end point, and just *pull them apart*, then the bits of string that you're left with in a straight line will give you the shortest route, and that process of pulling just gets you the solution very rapidly in a parallel computation, where all the others just hang by the wayside, so to speak.

This is an old idea, mentioned for example in (Dreyfus & Haugeland, 1974). A partly similar computer model of finding the shortest route in a network of roads could store at every junction a list of names of all other junctions, and for each other junction an indicating of which route goes to that target junction. Precomputing the best direction from node X to each of the other nodes, for every X would be worthwhile if the network is to he used often, without any searching. But the string version of the roadmap does not need such pre-computing.

5.5 In sum, we know surprisingly little about the kinds of problems that evolution solved, and the manner in which they were solved.

Now, I am not saying brains can build networks of string and pull them or anything like that; that's just an illustration of how, if you have the right representation, suitably implemented for a class of problems, then you can avoid combinatorially complex searches, which may grow exponentially with the number of components in your map, whereas with a special purpose device, e.g. a string map, the time it takes won't depend on how many strings you've got on the map; you just pull, and time required will depend only on the shortest route that exists in the network, even if that shortest route wasn't obvious from the original map.

That's a rather long-winded way of formulating the conjecture — or a roundabout way of supporting the conjecture — that there may be something about the way molecules perform computations using the combination of continuous change as things move through space, coming together, moving apart, folding, twisting, or whatever — and also sometimes snap into stable states that then resist disruption: as a result of quantum mechanisms, you can have stable molecular structures which are quite hard to separate using physical force. They may need catalytic processes to separate them, or extreme temperatures, or very strong forces, but they may nevertheless be able to move very rapidly in some conditions in order to perform computations.

Now there may be things about such mechanisms that enable searching for solutions to *certain* classes of problems to be done much more efficiently (by brains) than anything we could do with computers. As far as I know this is an open question.

So it *might* turn out that we need new kinds of technology in order to replicate the functions that animal brains perform, or it might not. I just don't know. I am not claiming that there's strong evidence for that; I am just saying that it might turn out that way, partly because I think we know less than many people think we know about what biological evolution achieved, and partly because of the many gaps that still remain between current AI systems and what humans and other animals can do.[9]

There are some other possibilities: we may discover that there are shortcuts no one has thought of so far, and then progress will occur much more quickly. I have an open mind; I would be very surprised, but it could happen.

[9] Some examples involving visual perception are assembled here, as challenges for vision researchers in AI, psychology and neuroscience: `http://www.cs.bham.ac.uk/research/projects/cogaff/misc/vision/`

6 Singularities of cognitive catch-up

6.1 What if it takes more than a lifetime to learn enough to make something new?

There is something that worries me much more than the singularity that most people talk about (the feared development of machines achieving human-level intelligence and perhaps taking over the planet or something). There's also what I call *singularities of cognitive catch-up*, SOCC, which I think we're close to, or may have already reached — at least in some areas of intelligence. I'll explain what I mean by that. (This is not a *unique* singularity, as will become clear.)

One of the products of biological evolution — and this is part of the answer to the earlier question about achievements of evolution — is that humans have not only the ability to make discoveries that none of their ancestors have ever made, but can also shorten the time required for similar achievements to be reached by their offspring, their descendants, and other humans. For instance, once we have worked out ways of doing complex computations, or ways of building houses, or ways of finding our way around,. . . , our children don't need to work the same things out for themselves by the same lengthy procedure; we can help them get there much faster.

In consequence, in the case of humans, it is not necessary for each generation to learn what previous generations learned *in the same way*. This may also be true to some extent of other species where the young learn from adults who have made novel discoveries, but humans also learn new ways to teach and new ways to learn, including use of verbal instructions, pictures, diagrams, models, and now computers. That ability to change the processes of learning as well as the content of what is learnt, makes a huge difference. So, once something has been learned we can enormously speed up the learning by new individuals. That has meant that the social processes supporting that kind of education of the young can enormously accelerate what would have taken perhaps thousands or millions of years for evolution to produce by encoding the relevant competences and knowledge in the genome.

But here's the catch: in many cases, in order for a new advance to happen — e.g. for something new to be discovered that wasn't known before, like Newtonian mechanics, or the theory of relativity, or Beethoven's musical style, or whatever — the individuals have to have traversed a significant amount of what their ancestors have learned, even if they do it much faster than their ancestors, to get to the point where they can see the gaps, discover new links, and make use of the possibilities for going further than their ancestors or contemporaries, have done. For example, it seems likely that Newton and Leibniz could not have developed calculus without building on the previous mapping of geometry into arithmetic by Descartes. If Beethoven had lived a thousand years earlier, ignorant of the work of Mozart, Haydn and others, he could not have composed his sonatas, symphonies, and string quartets. Even giants have to stand on the shoulders of giants.

In the case of knowledge of science, mathematics, philosophy, engineering and so on, there has been a lot of accumulated knowledge, growing faster and faster with time, increasing the amount of learning required in order to provide a new synthesis. Humans are living a *bit* longer than they used to, but the life span has not increased nearly as much as the amount of knowledge available to be learnt. So there may come a time when it is no longer possible for anyone living a normal human lifespan, to learn enough to understand the scope and limits of everything that has already been achieved in order to see the potential for going beyond it in a major new step.

That limit may be postponed by the use of new technology that speeds up learning processes, as search engines and other tools on the internet have done recently. But there could still be important discoveries, including discoveries relevant to understanding what evolution has achieved, that will not be made until some individual has assembled enough information (e.g. about diversity of life forms, about physics, chemistry, mathematics and computation) to discern a crucial new pattern. And that learning process may require two centuries – more than any human life span. Various forms of collaborative research, and new technologies, may reduce the time required, or may spread the discovery process over several generations of humans using a shared external memory. But it is simply possible that human information-processing capabilities are not up to the task of making important discoveries that in principle could be made, and thereafter all progress in advancing knowledge will be restricted to minor variants of what has gone before. (Research seminars and new publications often make me think we have already reached that stage.)

If we do reach that stage, we shall have reached a singularity of cognitive catch-up (SOCC). It is possible that the educational processes that enable individuals to learn faster than their ancestors did have reached their maximum speed and therefore only an extended lifetime, and perhaps also increased brain capacity, will allow the catching up required for certain major advances.

Unfortunately I now often encounter people presenting what *they* think of as new ideas which they've struggled to come up with, but who have not taken in relevant work done by other people, in other places at other times, including work that shows limitations or errors in the new ideas. That happens *despite* the availability of search engines, which now make it easier for people to get information. There just is too much information available and selecting the best bits is not easy, especially when the best available tools are distorted by requirements of financial sponsors.

When I was a student, finding out what other people had done in the field was a laborious process of going to the library, getting books, and so on, whereas now, I can often do things in seconds that would have taken hours. So, if only seconds are now needed for that kind of work, my lifespan has been effectively extended by a factor of several hundreds or thousands. But is the extension enough?

Technology *delays* the singularity, but it may not delay it enough. That's an open question; I don't know. Perhaps in some areas of knowledge, this is more of a problem than others. For instance, it may be that in some kinds of engineering, we're handing over more and more of the work to machines anyway, and they can go on doing it. Most of the production of computers now is done by a computer-controlled machinery, and even much software-writing. Although some of the design work is done by humans, a lot of *detail* of the design is done by machines that produce the next generation, which then produces the next generation, and so on.

This is not a singularity in which we are overtaken by potentially desirable machines. Rather it's a singularity in which some major advances, essential for our future survival, may depend on new kinds of machines that we cannot understand.

7 Spatial reasoning: a difficult problem

Okay, well, there are different branches of mathematics, and they have different properties. So, for instance, a lot of mathematics can be expressed in terms of logical structures or algebraic structures and those are pretty well suited for manipulation on computers, and if a problem can be specified using the logical/algebraic notation, and the solution method requires creating something in that sort of notation, then computers are pretty good.

There are many powerful mathematical tools available now, including theorem provers and proof checkers, and many others, which could not have existed fifty, sixty years ago, and they will continue getting better.

But there was something that I was alluding to earlier when I gave the example of how you can reason about what you will see by changing your position in relation to a door, where what you are doing is using your grasp of spatial structures; and how, as one spatial relationship changes, namely you come closer to the door or move sideways and parallel to the wall or whatever, then other spatial relationships change in parallel, so that the lines from your eyes through to other parts of the room on the other side of the doorway change: they spread out more as you go towards the doorway, but as you move sideways, they don't spread out differently, but focus on different parts of the environment: they access different parts of the other room.

Now, those are examples of ways of thinking about relationships and changing relationships which are not the same as thinking about what happens if I replace this symbol with that symbol, or if I substitute this expression in that expression in a logical formula. And at the moment, I do not believe that there is anything in AI amongst the mathematical reasoning community, the theorem-proving community, that can model the processes that go on when a young child starts learning to do Euclidean geometry and is taught things about space — for instance, I can give you a proof that the angles of any triangle add up to a straight line, 180 degrees.[10]

7.1 Example: Spatial proof that the angles of any triangle add up to a half-circle

There are standard proofs which involve starting with one triangle, then adding a line parallel to the base through the opposite vertex, then marking various pairs of angles equal, and drawing further consequences. Many years ago, a former student who had become a mathematics teacher, Mary Pardoe, came up with a lovely proof which I can demonstrate with this <holding up a pen>, and moving it around an imaginary triangle, as illustrated by the arrow in Figure 1.

Suppose I have a triangle and I put this pen on it, on one side (e.g. the bottom). I can then rotate it through one of the angles until it lies along the second side, and then I can rotate it again through the second angle, until it lies on the third side, and then finally rotate it through the third angle. After those three rotations it must end up on the *original* side, but it will have changed the direction it is pointing in, and it won't have crossed over itself, so it will have gone through a half-circle, and that says that the three angles of a triangle add up to the rotation of

[10] After giving this interview I learnt about a workshop in Edinburgh in 2012 on geometric theorem proving (Ida & Fleuriot, 2012). There has been very interesting progress, but it does not fill the gaps I am trying to describe.

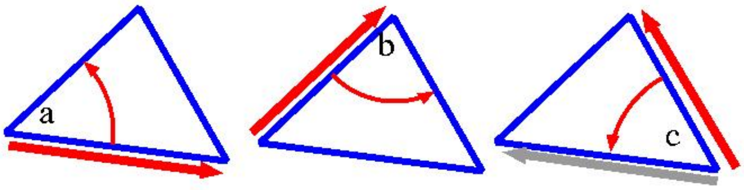

Figure 1: **Mary Pardoe's proof of the triangle sum theorem.**

half a circle, which is a beautiful kind of proof, and almost anyone can understand it.[11]

Some mathematicians don't like this proof, because they say it hides some of the assumptions, but nevertheless, as far as I am concerned, it is an example of a human ability to do reasoning which, once you've understood it, you can see will apply to any triangle.
Understanding that that is a proof requires understanding why the possibility of the process I have described does not depend on the location, size, orientation, colour, or even perfect depiction of the triangle. A mathematical learner has to grasp invariant features of a structure or process.

It is got to be a planar triangle, not a triangle on a globe for example, because on a globe the angles can add up to more than half a rotation; you can have three *right* angles if you have an equator, a line on the equator, a line going up to to the north pole, and then a right angle turn followed by another line going down to the equator. That produces a triangle on the surface of a sphere, with three right angles, adding up to more than a straight line.

But that's because the triangle is not in a plane, it is on a curved surface. In fact, that's one of the definitional differences you can take between planar and curved surfaces: how much the angles of a triangle add up to.

But our ability to *visualize* and notice the *generality* in that process, and see that you're going to be able to do the same thing using triangles that stretch in all sorts of ways, or if it's a million times as large, or if it's made of something different; if it's drawn in different colors or whatever — none of that's going to make any difference to the essence of that process. And that ability to see the commonality in a spatial structure which enables you to draw some conclusions with complete certainty. All of this is subject to the possibility of mistakes. Sometimes we make mistakes in mathematical reasoning, but when we make mistakes, we can normally discover them later, as happened in the history of geometrical theorem proving. Imre Lakatos wrote a wonderful book called *Proofs and Refutations*[12] — which I won't try to summarize — in which he presents examples: mistakes were made; for instance because mathematicians did not always realize there were subtle sub-cases which had slightly different properties, and they did not take account of that. But when such mistakes are noticed, they can often be rectified.

[11] See also `http://www.cs.bham.ac.uk/research/projects/cogaff/misc/triangle-sum.html`

[12] See `http://en.wikipedia.org/wiki/Proofs.and.Refutations`

7.2 Geometric results are fundamentally different from experimental results in chemistry or physics.

That's different from doing experiments in chemistry and physics, where you can't be sure results be the same on Mars or at a high temperature, or in a very strong magnetic field. With geometric reasoning, in some sense you've got the full information in front of you; even if you don't always notice an important part of it. That kind of geometrical reasoning, as far as I know, is not yet implemented anywhere in a computer. And most people who do research on trying to model mathematical reasoning, or research on human mathematical learning or development, ignore important issues, because of their own limited mathematical (and philosophical) education. They start with limited questions and assumptions because of how they were educated. I was taught Euclidean geometry at school including how to construct geometrical proofs. Were you?

(Adam ford: Yeah)

Many people are not now. Even mathematically promising children are taught set theory, and logic, and arithmetic, and algebra, rather than euclidean geometry. And so they don't use that bit of their brains, without which we wouldn't have buildings and cathedrals, and all sorts of things we now depend on.

There has been much research on automated geometrical theorem proving, but normally the axioms and theorems are translated into logical and algebraic formalisms, using Cartesian coordinate representations of geometry. So the machines prove theorems about sets of numbers and equations or inequalities relating numbers, not theorems about geometry such as Euclid proved, even if there's a strong structural relationship between the two domains.[13]

8 Is near-term artificial general intelligence (AGI) likely?

8.1 Two interpretations: a single mechanism for all problems, or many mechanisms unified in one program.

Well, this relates to what's meant by "general". When I first encountered the AGI community, I thought that what they all meant by *general* intelligence was *uniform* intelligence — intelligence based on some common simple (or maybe not simple but uniform) and powerful mechanisms for learning and inference. There are some people in the AGI community who are trying to produce things like that, often using ideas and techniques related to algorithmic information theory and compressibility of information, and so on. There have been interesting advances in this area, but usually tested only on problems to which the methods are well suited, not, for example, built into a model of learning and development in a child or other animal. An example is discussed in (Sloman, 2012).

But there's another sense of "general" which means that a system with general intelligence can do lots of different things, like perceive things, understand language, move around, make things, and so on — perhaps even enjoy a joke. That's something that's not nearly on the

[13] `http://www.cs.bham.ac.uk/research/projects/cogaff/misc/triangle-theorem.html`

horizon, as far as I know. Enjoying a joke isn't the same as being able to make laughing noises.

Nor is it the same as being able to recognize jokes, or generate jokes, as some AI programs can do, including amusing pun-generators, acronym generators, and others.[14]

So, there are these two notions of general intelligence (a) one that looks for one uniform, possibly simple, mechanism or collection of ideas and notations and algorithms, that will deal with any problem that's solvable, and (b) another that's general in the sense that it can do lots of different things, using a variety of forms of representation, learning mechanisms, reasoning mechanisms sensory mechanisms, motor control mechanisms, some innate and some learnt, that are all combined within an integrated architecture (which raises lots of questions about how you combine these things and make them work together). We humans are certainly of the second kind: we do all sorts of different things, using a wide variety of mechanisms that evolved at different times or are learnt at different stages, and other intelligent animals are also seem to be of the second kind, though none as general as humans.

It may turn out, in some near or remote future time — decades? or centuries? — that we shall have machines that are capable of solving any problem that is solvable, in a time that will depend on the nature of the problem, and they will be able to do it within a finite time. There are some problems that are solvable, but whose solution would require a larger universe and a longer history than the history of the universe (which is true of a chess player that explores the whole game tree), but apart from that constraint, these machines will be able to do anything (at least in principle) and they will have general intelligence.

I am ignoring the *real time* requirements of animal intelligence, needed for leaping through trees, escaping predators, catching prey, looking after errant offspring, and so on, all of which require perceptual processes, and many other processes, including planning, plan execution, motor control, all of which must work within relatively short time limits. Combining solutions to these problems with other mechanisms requires a multi-functional, but well integrated, information processing architecture.[15]

But to be able to do some of the kinds of things that humans can do — like the kinds of geometrical reasoning where you look at the shape and you abstract away from the precise angles and sizes and shapes and so on, and realize there's something general there (illustrated in Figure 1), as must have happened when our ancestors first made the discoveries that were eventually assembled in Euclidean geometry — may require mechanisms of a kind that we don't know about at present.

Maybe brains are using molecules and rearranging molecules in some way that supports that kind of reasoning. I am not saying they are: I don't know. I just don't see any simple or obvious way to map that kind of reasoning capability onto what we currently do on computers.

There is something I'll mention briefly before finishing: namely there is a kind of AI system that's sometimes thought of as a major step in that direction, namely we build a machine (or a software system) that can represent some physical or geometrical structure, and then be told about a change that's going to happen to it, and it can predict in great detail what will happen.

For instance, in some game engines, it is possible to say "We have all these blocks and other

[14] http://en.wikipedia.org/wiki/Computational_humor

[15] As investigated in the CogAff project. http://www.cs.bham.ac.uk/research/projects/cogaff/#overview

objects on the table and I'll drop one more block here", and then the program uses Newton's laws and properties of rigidity or elasticity or plasticity (etc.) and also facts about geometry and space and so on, to compute a very accurate representation of what will happen when the block lands on that specific pile of objects: it will bounce and go off, in a particular direction, and the others will all move in very specific ways that can be depicted by generating a video of the process. In that case, using more memory and more CPU power, you can increase the accuracy. The program can compute trajectories following a very wide variety of initial configurations, with more or less accuracy depending on the complexity and the duration of the predicted changes.[16]

But that ability to predict consequences is totally different from being able to look at *one* example, and work out what will happen in a whole *range* of cases at a higher level of abstraction, as in Pardoe's proof of the triangle sum theorem, illustrated above in Figure 1. The game engine does it in great detail for *just one* case, with just those precise things in that configuration, and it cannot discover new generalizations that would apply to other similar cases. So, in that sense (being able to make precise predictions about consequences of precisely specified events, and being able to solve a vast range of concrete problems) you may get AGI (artificial general intelligence) pretty soon, but it will be limited in what it can do. And the other kind of general intelligence which combines all sorts of different things, including human geometrical reasoning, and maybe other things, like the ability to find things funny, and to appreciate artistic features and other things may need forms of, types of, mechanism that we don't know about, and I have an open mind about how long those may take, or whether they can all be implemented on Turing machinery, or will require something more general.

9 Artificial General Intelligence (AGI) impacts

9.1 Implications of the two types of general intelligence.

As far as the first type of artificial general intelligence is concerned (the ability to make specific predictions about specific situations in a huge variety of cases), it could be useful for many kinds of applications — there are people who worry about whether a system that has that type of intelligence, might in some sense take over control of the planet. Well, humans often do stupid things, and the machines might do something stupid that would lead to disaster, but I think it is more likely that there would be other things done by humans that will lead to disaster — population problems, using up resources, destroying ecosystems, and whatever. But certainly it would go on being useful to have these calculating devices.

Now, as for the second kind of artificial general intelligence, combining many different capabilities in one system, I don't know. If we succeeded at putting together all the parts that we find in humans, we might just make an artificial human, and then we might have some of them as our friends, and some of them we might not like, and some of them might become teachers or whatever, ..., composers, etc.

But that raises a question: could they, in some sense, be superior to us, in their learning capabilities, their understanding of human nature, or maybe their wickedness or whatever?

[16] In the case of "chaotic" physical systems, or systems like pin-balls, arbitrarily small variations in initial conditions can have arbitrarily large effects after several bounces, and the results of predictions can vary widely depending on the numerical precision used – how many bits per number.

These are all issues on which I expect the best science fiction writers would give much better answers than anything I could do. But I did once fantasize when I wrote a book in 1978 *The Computer Revolution in Philosophy: Philosophy, science and models of mind*[17] in the Epilogue[18], that perhaps if we achieved that kind of thing, that they, the intelligent machines, would be wise, and gentle and kind, and realize that humans are an inferior species, but they have some good features, so they would keep us in some kind of secluded, restrictive kind of environment, and keep us away from dangerous weapons, and so on. And find ways of cohabiting with us. But that's just fantasy.

Adam: Awesome. Yeah, there's an interesting story *With Folded Hands* where the computers want to take care of us and want to reduce suffering and end up lobotomizing everybody, but keeping them alive so as to reduce the suffering.

Aaron: Not all that different from *Brave New World*, by Aldous Huxley, where it was done with drugs and so on, but different humans are given different roles in that system.

There's also *The Time Machine*, by H.G. Wells, where, in the distant future, humans have split/evolved into two types: The Eloi lived on the surface of the earth, and the Morlocks lived underground, and were intelligent and unattractive. The Eloi lived on the surface of the planet. They were pleasant and pretty but not very bright, and they were fed on by the Morlocks!

Adam: Yeah ... that's strange, ... in the future.

Aaron: As I was saying, if you ask science fiction writers, you'll probably come up with a wide variety of interesting answers.

Adam: I certainly have; I have spoken to Kim Stanley Robinson, Sean Williams, David Brin, and, ...

Aaron: Did you ever read a story by E.M. Forster called *The Machine Stops*[19] It's a very short story, written around 1909, about a future time when people sit in their rooms, in front of screens, communicating with others by typing at a terminal. They don't meet. They have debates, and give lectures to their audiences online. One of them is a woman whose son says "I'd like to see you" to which she responds something like "What's the point? You can already talk to me", but he wants to come and talk to her face to face. I won't tell you how it ends, but....

Adam: Reminds me of the Internet.

Aaron: Well, yes; he invented it, over sixty years in advance! It was quite extraordinary that he was able to do that, before most of the components that we now need for it existed.

Adam: Another person who did that was Vernor Vinge, in a novella called *True Names*.

Aaron: When was that written? (I have not read it.)

Adam: The seventies.

Aaron: A lot of the technology was already around then. The original bits of internet were already working by about 1973. In 1974, after having spent a year learning about AI and programming in Edinburgh, I was back in Sussex University trying to learn LOGO, the programming language, in order to decide whether it was going to be useful for teaching AI. Max Clowes kindly gave me access to his teletype machine, in Brighton. There was no screen, only paper coming out with printing on it: what I typed and what a machine in California typed in response. My machine transmitted ten characters a second from Sussex University

[17] Online at `http://www.cs.bham.ac.uk/research/projects/cogaff/crp/`

[18] See `http://www.cs.bham.ac.uk/research/projects/cogaff/crp/epilogue.html`

[19] Available online: `http://archive.ncsa.illinois.edu/prajlich/forster.html`

to University College London computer lab by telegraph cable, from there to somewhere in Norway via another cable, from there by satellite to California and eventually to a computer in Xerox Palo Alto Research Center (Xerox PARC), where they had implemented a computer with a LOGO system on it. I also had online conversations with someone I had met previously in Edinburgh, Danny Bobrow, who had kindly arranged for me to have access to the Xerox machine.

So there I was typing into a machine thousands of miles away. Furthermore, it was duplex typing, so every character I typed didn't show up on my terminal until it had gone all the way to California and been echoed back. The characters would come back about four seconds later after each keypress.

That was already the Internet, though in a very primitive state, and I suspect Vernor Vinge wrote after that kind of thing had already started.

I mentioned H.G. Wells' *The Time Machine* earlier. I recently discovered, because David Lodge had written a sort of semi-novel about Wells *A Man of Parts*[20] that Wells had invented Wikipedia long in advance: he had the notion of an encyclopedia that was free to everybody, to which everybody could contribute, in a collaborative effort.

So, go to the science fiction writers to find out the future, or a range of possible futures.

Adam: Well, the thing is with science fiction writers, they have to maintain some sort of interest in their readers, after all the science fiction which reaches us is the stuff that publishers want to sell, and so there's a little bit of a bias towards making a plot device there, and so the dramatic sort of appeals to our amygdala, our lizard brain; will sort of be there obviously, will be mixed in. But I think that they do come up with some amazing ideas. I think it's worth trying to make these predictions: we should focus more time on strategic forecasting, I mean take that seriously.

Aaron: Well, I am happy to leave that to others; I just want to try to understand these problems that bother me, about how things work. And it may be that some would say that's irresponsible if I don't think about what the implications will be. Well, understanding how humans work *might* enable us to make surrogate humans — but I suspect it wont happen in this century; I think it's going to be too difficult.

10 Further Reading on the Meta-Morphogenesis project

Here are some links to online documents presenting various aspects of the Meta-Morphogenesis project, including links to further sources. The ideas are developing and changing rapidly.

- The Meta-Morphogenesis Project – How a Planet can produce Minds, Mathematics and Music. `http://www.cs.bham.ac.uk/research/projects/cogaff/misc/meta-morphogenesis.html`
- Adam Ford's video recording of the tutorial on the Meta-Morphogenesis Project, given at the AGI 2012 conference, Oxford. `http://www.youtube.com/watch?v=BNul52kFI74`
- Meta-Morphogenesis and Toddler Theorems: Case Studies. `http://www.cs.bham.ac.uk/research/projects/cogaff/misc/toddler-theorems.html`

[20] `http://www.amazon.co.uk/A-Man-Parts-David-Lodge/dp/0099556081`

- A draft, changing, list of types of transitions in biological information-processing. http://www.cs.bham.ac.uk/research/projects/cogaff/misc/evolution-info-transitions.html
- Biology, Mathematics, Philosophy, and Evolution of Information Processing. http://www.cs.bham.ac.uk/research/projects/cogaff/misc/bio-math-phil.html

Acknowledgements

Without the work done by Dylan Holmes producing the first transcript of the video interview, this document would not have been possible.

References

Cooper, S. B., & Leeuwen, J. van (Eds.). (2013). *Alan Turing: His Work and Impact*. Amsterdam: Elsevier.

Craik, K. (1943). *The nature of explanation*. London, New York: Cambridge University Press.

Dreyfus, H., & Haugeland, J. (1974). The computer as a mistaken model of the mind. In S. Brown (Ed.), *Philosophy of Psychology* (pp. 247–258.). London: Macmillan.

Gibson, J. (1966). *The Senses Considered as Perceptual Systems*. Boston: Houghton Mifflin.

Gibson, J. J. (1979). *The ecological approach to visual perception*. Boston, MA: Houghton Mifflin.

Ida, T., & Fleuriot, J. (Eds.). (2012, September). *Proc. 9th Int. Workshop on Automated Deduction in Geometry (ADG 2012)*. Edinburgh: University of Edinburgh. Available from http://dream.inf.ed.ac.uk/events/adg2012/uploads/proceedings/ADG2012-proceedings.pdf

Jablonka, E., & Lamb, M. J. (2005). *Evolution in Four Dimensions: Genetic, Epigenetic, Behavioral, and Symbolic Variation in the History of Life*. Cambridge MA: MIT Press.

Marr, D. (1982). *Vision*. San Francisco: W.H.Freeman.

Sloman, A. (1971). Interactions between philosophy and AI: The role of intuition and non-logical reasoning in intelligence. In *Proc 2nd ijcai* (pp. 209–226). London: William Kaufmann. Available from http://www.cs.bham.ac.uk/research/cogaff/04.html#200407

Sloman, A. (1978). What About Their Internal Languages? Commentary on three articles by Premack, D., Woodruff, G., by Griffin, D.R., and by Savage-Rumbaugh, E.S., Rumbaugh, D.R., Boysen, S. in *Behavioral and Brain Sciences* Journal 1978, 1 (4). *Behavioral and Brain Sciences*, *1*(4), 515. (http://www.cs.bham.ac.uk/research/projects/cogaff/07.html#713)

Sloman, A. (1979). The primacy of non-communicative language. In M. MacCafferty & K. Gray (Eds.), *The analysis of Meaning: Informatics 5 Proceedings ASLIB/BCS Conference, Oxford, March 1979* (pp. 1–15). London: Aslib. Available from `http://www.cs.bham.ac.uk/research/projects/cogaff/81-95.html#43`

Sloman, A. (2008). *Evolution of minds and languages. What evolved first and develops first in children: Languages for communicating, or languages for thinking (Generalised Languages: GLs)?* (Research Note No. COSY-PR-0702). Birmingham, UK. Available from `http://www.cs.bham.ac.uk/research/projects/cosy/papers/#pr0702`

Sloman, A. (2011, Sep). *What's vision for, and how does it work? From Marr (and earlier) to Gibson and Beyond.* Available from `http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#talk93` (Online tutorial presentation, also at `http://www.slideshare.net/asloman/`)

Sloman, A. (2012). Chapter 4A. Aaron Sloman on Schmidhuber's "New Millennium AI and the Convergence of History 2012". In A. H. Eden, J. H. Moor, J. H. Soraker, & E. Steinhart (Eds.), *The Singularity Hypotheses: A Scientific and Philosophical Assessment* (pp. 79–80.). Springer, Berlin Heidelberg. Available from `http://www.cs.bham.ac.uk/research/projects/cogaff/12.html#1208`

Sloman, A. (2013). Virtual machinery and evolution of mind (part 3) meta-morphogenesis: Evolution of information-processing machinery. In S. B. Cooper & J. van Leeuwen (Eds.), *Alan Turing - His Work and Impact* (p. 849-856). Amsterdam: Elsevier. Available from `http://www.cs.bham.ac.uk/research/projects/cogaff/11.html#1106d`

Sloman, A., & Chappell, J. (2007). Computational Cognitive Epigenetics (Commentary on (Jablonka & Lamb, 2005)). *Behavioral and Brain Sciences*, *30*(4), 375–6. (http://www.cs.bham.ac.uk/research/projects/cosy/papers/#tr0703)

Turing, A. M. (1936). On computable numbers, with an application to the Entscheidungsproblem. *Proc. London Math. Soc.*, *42*(2), 230–265. Available from `http://www.thocp.net/biographies/papers/turing_oncomputablenumbers_1936.pdf`

Turing, A. M. (1952). The Chemical Basis Of Morphogenesis. *Phil. Trans. R. Soc. London B 237*, *237*, 37–72.